Structural Equation Models for Finite Mixtures – Simulation Results and Empirical Applications

Dirk Temme[1], John Williams[2] and Lutz Hildebrandt[3]

[1] Institute of Marketing, Humboldt University of Berlin, 10178 Berlin, Germany
[2] Department of Marketing, Otago University, Dunedin, New Zealand
[3] Institute of Marketing, Humboldt University of Berlin, 10178 Berlin, Germany

Abstract. Unobserved heterogeneity is a serious but often neglected problem in structural equation modelling (SEM) challenging the validity of many empirical results. Recently, a finite mixture approach to SEM has been proposed to resolve this problem but until now only a few studies analyse the performance of the relevant software. The contribution of this paper is twofold: First, results from a Monte Carlo study into the properties of the program system MECOSA are presented. Second, an empirical application to data from a large-scale consumer survey in the fast moving consumer goods industry is described.

Keywords. Structural equation modelling, Unobserved heterogeneity, Model-based clustering, Finite mixtures, Monte Carlo simulation

1 Introduction

Structural equation modelling (SEM) is an established, widely used methodology in the social sciences (e.g., sociology, psychology, marketing). One of its key benefits is that it enables the estimation of relationships between unobservable theoretical constructs (e.g., attitudes, customer satisfaction) which are operationalised by multiple observed, albeit individually imperfect measures.

Empirical applications of SEM typically rest on the assumption that either the analysed sample is homogenous with respect to the underlying model or that heterogeneity can be adequately taken into account by forming sub-samples using one or two observed criteria (e.g., gender, age, brand loyalty). But, if unobserved heterogeneity is (still) substantial, parameter estimates might be seriously biased (Jedidi, Jagpal and DeSarbo, 1997). As an ad hoc solution to this problem one might follow a two-step procedure, which combines cluster analysis in the first step with a multi-group analysis in the second step. However, both theoretical considerations as well as simulation evidence (Görz, Hildebrandt and Annacker, 2000; Jedidi, Jagpal and DeSarbo, 1997) have shown that this procedure is inappropriate.

Meanwhile two alternative solutions to the problem of unobserved heterogeneity in SEM have been proposed. Whereas the finite mixture approach presumes that heterogeneity in the population can be sufficiently captured by a limited number of homogenous segments, hierarchical Bayes methods (Ansari, Jedidi and Jagpal, 2000) rest upon the idea that individual-level parameters follow a continuous heterogeneity

distribution. In this paper, we exclusively focus on finite mixture SEM. Since until now only a few studies exist which analyse the performance of this methodology we first describe the design and main results of a Monte Carlo simulation study into the properties of the MECOSA approach to conditional finite mixture SEM. Second, the results of an empirical application of MECOSA to data from a large-scale consumer survey on attitudes towards a specific brand in the German fast moving consumer goods industry are reported.

2 Finite Mixture Structural Equation Models

2.1 Conditional and Unconditional Models

The finite mixture approach assumes that the sample is a discrete mixture of a limited, but generally unknown number of components each characterised by a specific distribution. With respect to the distributional assumptions, two types of finite mixture SEM can be distinguished. Unconditional models (Yung, 1997; Jedidi, Jagpal and DeSarbo, 1997; Dolan and van der Maas, 1998) rely on the assumption that the endogenous and exogenous variables follow a multivariate normal distribution within the different components. In contrast, for conditional models the somewhat weaker, more realistic assumption applies that the dependent variables are normally distributed given some exogenous regressor variables (e.g., demographics). To the best of our knowledge only two commercial software programs for the estimation of conditional finite mixture SEM are presently available (MECOSA, Arminger, Wittenberg and Schepers, 1996; Mplus, Muthén and Muthén, 1998).

2.2 The MECOSA Approach

The following description of the MECOSA model for conditional multivariate normal mixtures is mainly based on Arminger, Stein and Wittenberg (1999; abbreviated as ASW below). Let $\mathbf{y}_i, i = 1,2,\ldots,n,$ be a p-dimensional vector of continuous dependent random variables and $\mathbf{x}_i$ a q-dimensional vector of continuous or dummy-type independent variables. The sample points $(\mathbf{y}_i, \mathbf{x}_i)$ are i.i.d. with density $h(\mathbf{y}_i, \mathbf{x}_i) = f(\mathbf{y}_i \mid \mathbf{x}_i) \cdot g(\mathbf{x}_i)$, where $g(\mathbf{x}_i)$ is the marginal density of the exogenous variables. If the dependent variables $\mathbf{y}_i$ are multivariate normal in each component conditional on the regressors $\mathbf{x}_i$, the conditional density is given by the following mixture:

$$f(\mathbf{y}_i \mid \mathbf{x}_i) = \pi_1 \phi(\mathbf{y}_i; \mu_{i1}; \Sigma_1) + \pi_2 \phi(\mathbf{y}_i; \mu_{i2}; \Sigma_2) + \cdots + \pi_K \phi(\mathbf{y}_i; \mu_{iK}; \Sigma_K),$$

where $\pi_k, k = 1,2,\ldots,K,$ are the mixing proportions of the K mixture components subject to the following constraints: $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. $\phi(\bullet; \mu_{ik}; \Sigma_k)$ denotes the multivariate normal density with mean vector μ_{ik} and covariance matrix Σ_k. The conditional mean $E(\mathbf{y}_i \mid \mathbf{x}_i, k)$ is specified as a reduced form multivariate linear model

$$\mu_{ik} = \gamma_k + \Pi_k \mathbf{x}_i ,$$

where γ_k is a p-dimensional vector of regression constants and Π_k is a $p \times q$ matrix of regression coefficients. The conditional covariance matrix Σ_k contains the variances and covariances of the regression residuals.

The conditional means and covariances are parameterised by component-specific mean- and covariance structure models, for example a conditional LISREL model, where the free parameters are collected in a vector ϑ:

$$\eta_i \mid (\mathbf{x}_i, k) = \mathbf{B}_k \eta_i + \Gamma_k \mathbf{x}_i + \zeta_i^{(k)},$$
$$\mathbf{y}_i = \nu_k + \Lambda_k \eta_i^{(k)}{}_i + \delta_i^{(k)}.$$

For the conditional mean of the endogenous variables this implies

$$E(\mathbf{y}_i \mid \mathbf{x}_i, k) = \nu_k + \Lambda_k (\mathbf{I} - \mathbf{B}_k)^{-1} \Gamma_k \mathbf{x}_i = \gamma_k + \Pi_k \mathbf{x}_i ,$$

where $\gamma_k = \nu_k$ and $\Pi_k = \Lambda_k (\mathbf{I} - \mathbf{B}_k)^{-1} \Gamma_k$. The conditional covariance matrix is specified as

$$V(\mathbf{y}_i \mid \mathbf{x}_i, k) = \Lambda_k (\mathbf{I} - \mathbf{B}_k)^{-1} \Psi_k (\mathbf{I} - \mathbf{B}_k)^{-1'} \Lambda_k' + \Theta_k = \Sigma_k .$$

MECOSA currently offers three estimation methods: Minimum Distance Estimation (MDE), Direct EM (EM) and EM Gradient (EMG). The MDE method is a two-step procedure. In the first stage, an unrestricted multivariate regression model for finite mixtures is estimated by an EM algorithm. In the second stage, a minimum distance estimator is used to estimate the group-specific fundamental parameters ϑ based on the reduced form parameter estimates $\hat{\gamma}_k, \hat{\Pi}_k$ and $\hat{\Sigma}_k$ in step one. In contrast, maximisation in the M-Step of both direct EM algorithms (EM and EMG) is directly performed with respect to the fundamental parameters and the mixing proportions π_k.

3 Simulation Study

3.1 Experimental Design

The experimental design of our simulation study (for a more comprehensive description see Williams, 2002; Williams, Temme and Hildebrandt, 2002) designed to test the performance of the MECOSA approach to the estimation of finite mixture SEM extends the experimental factors: 1. Distribution of variables (skewed or normal) and 2. Estimation method (MDE, EM and EM Gradient) used in the study by ASW by four additional factors: 3. Number of groups (two or three), 4. Group proportions (equal or mixed), 5. Knowledge about number of groups (known or unknown) and 6. Differences between the parameters in each group (close or far). In total this design leads to 96 different conditions. For each condition 500 valid data sets with 2000 observations each have been simulated with GAUSS according to a parameterised structural equation model and subsequently analysed using MECOSA.

3.2 Main Results

For all conditions with an unknown number of components we used the ad hoc procedure (a modified likelihood ratio test) implemented in MECOSA to estimate the number of groups. Whereas for two groups the number of components was almost always estimated correctly under all conditions, a completely different picture emerged for those three group conditions with close parameters and mixed proportions. More than 1000 data sets had to be simulated to achieve 500 replications with the number of groups estimated correctly, which is a prerequisite for meaningful comparisons between the results for different conditions.

The performance of MECOSA for the two group conditions in terms of estimator variance and bias is shown in Table 3.1. Both the highest mean MAD (mean absolute deviation) and the highest mean bias result for the MDE under conditions of mixed proportions and close parameters. Further analysis of the individual parameter recovery using a balanced ANOVA shows that the largest effect on parameter recovery is due to the differences in parameter values between groups (far versus close parameters). In addition, the negative effect of mixed proportions on parameter recovery is quite pronounced, especially in conjunction with close parameters. Under most conditions the choice of the estimation method seems to have only a negligible effect but for mixed proportions combined with close parameters MDE performs considerably worse than Direct EM and EM Gradient.

Table 3.1. Mean MAD and Mean Bias for the two group conditions

		Mean MAD				Mean Bias			
		Known		Unknown		Known		Unknown	
		Equal	Mixed	Equal	Mixed	Equal	Mixed	Equal	Mixed
		Close Parameters							
Skewed	MDE	.0339	.0443	.0338	.0452	-.0049	-.0116	-.0049	-.0127
	EM	.0316	.0409	.0319	.0400	-.0010	-.0018	-.0014	-.0012
	EMG	.0317	.0399	.0317	.0392	-.0021	-.0018	-.0013	-.0020
Normal	MDE	.0337	.0456	.0340	.0451	-.0047	-.0134	-.0061	-.0118
	EM	.0320	.0400	.0322	.0396	-.0015	-.0020	-.0002	-.0016
	EMG	.0323	.0398	.0319	.0402	-.0018	-.0019	-.0029	-.0033
		Far Parameters							
Skewed	MDE	.0230	.0283	.0227	.0282	-.0016	-.0033	-.0022	-.0028
	EM	.0221	.0272	.0226	.0268	-.0003	-.0004	-.0002	-.0007
	EMG	.0223	.0270	.0224	.0272	-.0004	-.0006	-.0007	-.0006
Normal	MDE	.0222	.0269	.0223	.0271	-.0018	-.0026	-.0019	-.0024
	EM	.0216	.0265	.0216	.0259	-.0007	-.0010	-.0004	-.0004
	EMG	.0215	.0265	.0217	.0257	-.0002	-.0005	-.0005	-.0006

4 Empirical Application

For marketing practice measuring customer based brand equity is of great importance in order to assess the effect of long-term investments in a brand. Marketing research companies and academics alike have proposed various approaches to measure brand equity. In this study we focus on the Brand Potential Index® (BPI) developed by the German company GfK, Nuremberg. The BPI consists of 9 items

which form a unidimensional confirmatory factor model. The following items are measured on a 7-point scale with the endpoints (7) "I totally agree" and (1) "I totally disagree" (items 2 – 9 are all formulated in comparison to other brands): 1. will buy the brand in the future (*buy*), 2. will pay more for the brand (*pay*), 3. identify with the brand (*identify*), 4. trust the brand (*trust*), 5. will recommend the brand (*recom*), 6. brand positively differentiates itself (*positive*), 7. like the brand (*like*) 8. regret if brand is not available (*regret*) and 9. brand is of higher quality (*quality*). Data for this study has been collected by a consumer survey of 1048 subjects who rated several competing brands in a specific convenience food category on the 9 BPI items. In addition, socio-demographic (e.g., age, household size) and behavioural (brand loyalty) information on the respondents has been gathered. These variables were used as exogenous regressors in a conditional confirmatory factor model. After eliminating several outliers with inconsistent answers a sample size of 1037 remained.

The data has been analysed with MECOSA, using first the two-stage MDE procedure. Since there was no clear a priori information about the number of segments, we let MECOSA estimate the number of components in the data. Based on the finite mixture multivariate regression of the BPI items on the exogenous regressors the ad hoc test ($LR_{ad\ hoc}$) as well as the parametric bootstrap test (LR_{Boot}) point to a three group solution although the BIC indicates a two group segmentation (see Table 4.1.) Since the entropy measure (Ramaswamy, DeSarbo, Reibstein and Robinson, 1993) for the three group estimation is .791, which indicates a good separation of the groups, we decided in favour of three groups.

Table 4.1. Summary statistics for model selection results

K	LL	$LR_{ad\ hoc}$	df	p	LR_{Boot}[1]	BIC
1	-13655.64	-	-	-	-	28248.73
2	-12932.54	1446.20	270	.000	331.15	27746.93
3	-12620.46	624.16	405	.000	450.29	28067.17
4	-12407.62	425.68	540	.999	450.47	28585.88

[1]Greatest Boot-strap LR value

With respect to the unconditional means of the BPI items (not shown here for space reasons) a clear order of the three groups emerges. Group 1 (N = 181) has the highest values (except for the item *buy* where the mean in group 2 is higher), followed closely by group 2 (N = 531). Group 3 (N = 325) clearly differentiates itself from the two other groups since all item means are considerably lower. For most of the regressors the mean differences between the groups are only minor. The most pronounced difference concerns the variable *loyalty*. Whereas in the first group about 71% of the subjects report that most of the time they buy the focal brand for the two other groups this percentage is considerably lower (63% for group 2 and only 53% for group 3). Belonging to the western or eastern parts of Germany also makes a difference: In group 3 23% of the subjects live in East Germany whereas in group 1 this holds for only 17%. By far the strongest effect on the BPI items occurs for the regressor *loyalty*, which is, as can be expected, positive and highly significant in all three groups. Only for group 3 *age* positively

influences the items *positive* and *regret*. Respondents from those German federal states where the analysed food category is especially popular tend to react more positively to the items *recom* and *like* than inhabitants of other states.

In the next stage, we used the three alternative estimation methods to estimate the fundamental parameters of a conditional common regression model (CRM; Yung, 1997):

$$\mu_{ik} = \nu_k + \Lambda\alpha_k + \Lambda\Gamma_k \mathbf{x}_i,$$

$$\Sigma_k = \Lambda\Psi_k\Lambda^{'} + \Theta_k.$$

The estimation using MDE yields a χ^2 statistic of 606.76 with 325 degrees of freedom. Although the model is rejected by this test (p = .000), the RMSEA = .042 is below the cut-off value of .05 and thus indicates a good fit. Because of space limitations we only report those parameter estimates where significant differences between the three estimation methods occur.

Table 4.2. Parameter estimates for the conditional common regression model

	MDE			EM			EMG		
	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
ν_1	**2.21**	**2.67**	**2.87**	**2.49**	**2.98**	**3.18**	**2.49**	**2.98**	**3.17**
ν_2	**1.92**	.93	1.20	**2.45**	**1.62**	**1.42**	**2.45**	**1.61**	**1.43**
ν_3	**1.64**			**2.15**			**2.15**		
ν_4	**2.14**			**2.42**			**2.42**		
ν_5	**2.15**			**2.51**			**2.51**		
ν_6	**2.23**			**2.64**			**2.64**		
ν_7	**1.90**			**2.42**			**2.42**		
ν_8	**1.85**			**2.33**			**2.33**		
ν_9	**2.41**			**2.73**			**2.74**		
α	.00[a]	**2.50**	.76	.00[a]	**1.79**	.41	.00[a]	**1.81**	.40
γ_1	.23	-.10	.09	.22	-.12	**.27**	.23	-.13	**.27**
γ_2	-.12	-.06	.08	**-.33**	-.03	.05	**-.34**	-.02	.05
γ_3	.25	.09	-.25	**.44**	.12	-.23	**.43**	.11	-.22
γ_4	**.09**	-.01	.02	.07	-.00	.01	.07	-.00	.01
γ_5	.16	.03	.09	**.45**	-.01	.15	**.47**	-.03	.16
γ_6	.14	-.03	.21	-.37	-.06	.09	-.37	-.05	.09
γ_7	**-1.20**	.09	.17	-.49	-.08	.19	-.52	-.07	.18
γ_8	-.21	**.23**	.21	.18	**.30**	.13	.18	**.29**	.15
γ_9	**2.68**	**.90**	**.70**	**2.87**	**1.08**	**.91**	**2.86**	**1.08**	**.92**

[a]fixed; significant parameters (α = .05) are set in bold type

Whereas the direct EM and the EMG methods yielded almost identical estimates, the MDE estimations in part considerably deviate from those of the EM methods (see Table 4.2.), leading to substantially different conclusions. For example, whereas *age* (γ_1) has a significant positive effect on BPI in group 3 when estimated using the EM methods (see also the results for the first step) this parameter is insignificant in the MDE solution. Also the MDE estimate for the dummy variable

West/East Germany (γ_7) seems unreasonably high. Overall, in line with the simulation results, the EM estimates for this three groups/mixed proportions case seem to be more trustworthy than the MDE estimates.

5 Conclusion

Since both the simulation results and the empirical application have shown that in situations which typically occur in empirical studies (mixed proportions and/or close parameters) the MDE method performs worse than the EM algorithms, we suggest to use only the first stage of the MDE procedure and to estimate the fundamental parameters by using the EM Gradient method.

References

Ansari, A., Jedidi, K. & Jagpal, S. (2000). A hierarchical bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science*, **19 (4)**, 328-347.

Arminger, G., Stein, P. & Wittenberg, J. (1999). Mixtures of conditional mean- and covariance structure models. *Psychometrika*, **64**, 475-494.

Arminger, G., Wittenberg, J. & Schepers, A. (1996). *MECOSA 3 User Guide*. Friedrichsdorf, Germany: Additive.

Görz, N., Hildebrandt, L. & Annacker, D. (2000). Analyzing multigroup data with structural equation models. In: *Proceedings of the 23rd Annual Conference of the GfKl*, 312-319. Berlin: Springer.

Dolan, C.V. & van der Maas, H.L.J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika*, **63**, 227-253.

Jedidi, K., Jagpal, H. & DeSarbo, W. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, **16 (1)**, 39-59.

Muthén, L.K. & Muthén, B.O. (1998). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Ramaswamy, V., DeSarbo, W.S., Reibstein, D.J. & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketing Science*, **12 (1)**, 103-124.

Williams, J. (2002). *Mean and covariance structure models for finite mixtures*. Unpublished Ph.D. thesis, Otago University, Department of Marketing, Dunedin, New Zealand.

Williams, J, Temme, D. & Hildebrandt, L. (2002). A Monte Carlo study of structural equation models for finite mixtures. *Discussion Paper SFB 373*. Berlin: Humboldt Universität zu Berlin.

Yung, Y.F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, **62**, 297-330.